Sampling and local algorithms in large graphs

Endre Csóka

Doctoral School: Mathematics
Director: Miklós Laczkovich, member of the Hungarian Academy of Sciences

Doctoral Program: Pure Mathematics
Director: András Szűcs, member of the Hungarian Academy of Sciences

Supervisor:
László Lovász, member of the Hungarian Academy of Sciences

Department of Computer Science, Eötvös Loránd University
December 2012
Very large graphs are present in almost all areas of the world. These appear in biological systems, e.g. the brain; in physics, e.g. the graph of the bonds between the molecules of a solid; furthermore, the internet, the traffic system, the electrical grid and social networks like the acquaintance graph of all people are also important very large graphs. In many cases, these graphs are not only huge but it is hopeless to get to know them precisely. However, we still have a chance to get to know some important properties of them.

In the beginning, the statistical analysis of very large graphs became popular in other areas of science, especially in statistical physics. They measured and measure the degree distribution of the graph, sometimes together with the correlation of the degrees or the density of triangles, and some other „local” data. Then the conclusions are made from generating large random graphs with these parameters, and measuring the properties of these graphs. They usually use heuristic algorithms for generating random graphs, which do not guarantee uniform randomness at all. However, scientists of these areas are very satisfied with the results. Understanding the background of this phenomenon was the main motivation for the mathematical theory of very large graphs.

The mathematical description of the question is, which graph properties and parameters can be estimated by a constant-size sampling. A graph parameter is estimable if for each $\varepsilon > 0$, there exists a constant-time sampling algorithm such that for each graph, this returns a value with an error at most $\varepsilon$ from the parameter value of the graph, in expectation. There are two different models for sampling algorithms.

The definition presented by Oded Goldreich, Shafi Goldwasser and Dana Ron [18] is the following. We take a constant number of vertices uniformly at random (allowing multiplicities), and we take the induced subgraph on these nodes. This is the simplified but equivalently strong version of the definition that we can use the following two kind of steps in a constant number of times. One kind of step is choosing a uniform random node and the other one is asking about two nodes whether there is an edge between them. The limit theory for this sampling method was developed by László Lovász with Balázs Szegedy [27], and with Borgs, Chayes, T. Sós and Vesztergombi [7], [6]. This theory could answer the main questions like what the testable properties and parameters are, therefore, this topic is complete in some sense. We call it the theory of dense graphs.

While this topic itself is complete, it has connections with several other topics, and there is further research in these directions. The language it uses already turned out to be useful for some existing and new topics. As an example, in Chapter ??, we show a new conjecture and some partial results on it. But returning to the original motivation, this theory is useful only for dense graphs, that is graphs with $\Theta(n^2)$ edges; but unfortunately, most real-life graphs are not so dense at all, therefore, the sampling method returns with an empty graph for almost sure.

However, there is another model by Oded Goldreich and Dana Ron [19], with the limit theory developed by Itai Benjamini and Oded Schramm [4]. It deals with bounded degree graphs (or, at least, graphs with $O(n)$ number of edges). This model fits much better to the real-life networks, but its mathematical theory turned out to be a much more difficult task. While there are several important results about it, this theory is far from being completed. Moreover, this includes algorithmically undecidable questions, as we will show it in Chapter ???. The larger part of my dissertation is about this theory, called the theory of sparse graphs.

Here, sampling means the following. We choose a constant number of vertices uniformly at random, and we take the constant-radius neighborhood of each. This is the simplified but equivalently strong version of the definition that we can use the following two kind of steps in a constant number of times. One kind of step is choosing a uniform random node and the other one is getting the list of neighbors of a node.

There are some other topics about sampling from large structures, such as permutations.
by Kohayakawa [21], partially ordered sets by Janson [23], abelian groups by Szegedy [33] and metric spaces by Gromov [20] and Elek [16]. As the theory of dense graphs is the first and the only complete theory, therefore, this provides useful observations and suggestions for the other theories. We show an overview of these results and its connections to the theory of sparse graphs.

Homomorphism numbers are a common tool for the two models. For two graphs $F$ and $G$, the homomorphism number $\hom(F, G)$ is the number of edge-preserving mappings $h: V(F) \to V(G)$. Formally,

$$\hom(F, G) = \left| \left\{ h: V(F) \to V(G) \mid \forall (x, y) \in E(F): (h(x), h(y)) \in E(G) \right\} \right|.$$  

In the theory of dense graphs and in the theory of sparse graphs, sampling a graph $G$ expresses the similarity of graphs with respect to the sampling. We define $\tilde{t}$ expressed by getting approximate values for $t(F, G) = \hom(F, G)/|V(G)|^{\text{W}(F)}$ and $\text{hom}(F, G)/|V(G)|$, respectively, for a bounded number of graphs $F$. That is, the only difference is the way of normalizing.

Consider the space $\mathcal{G}$ of all isomorphism types of graphs. We put a topology $\mathcal{T}$ on it, to express the similarity of graphs with respect to the sampling. We define $\mathcal{T}$ as the coarsest topology in which the homomorphism densities $t(F, \cdot)$ are continuous for all graphs $F$. In other words, a sequence of graphs $G_1, G_2, \ldots$ is convergent in $\mathcal{T}$ if for all graphs $F$, the sequence $t(F, G_n)$ is convergent.

Notice that if we multiply all nodes of a graph by the same number, then these two graphs are equivalent in this topology, therefore, we do not distinguish them. On the other hand, graphs which do not arise from the same graph by node-multiplication of the same graph, are equivalent in this topology, therefore, we do not distinguish them. On the other hand, graphs which do not arise from the same graph by node-multiplication of the same graph, are not equivalent in the topology.

Symmetric measurable functions $[0, 1]^2 \to [0, 1]$ are called graphons. Graphs on the set of points $\{0, 1, \ldots, n-1\}$ are represented by the graphon defined by

$$w(x, y) = \begin{cases} 1 & \text{if there is an edge between } \lfloor nx \rfloor \text{ and } \lfloor ny \rfloor \\ 0 & \text{otherwise.} \end{cases}$$

Sampling from a graphon means that we take a constant number of uniform random values from $[0, 1]$, we take the submatrix according to the rows and columns at these values, and we take the graph defined by this matrix as adjacency matrix. Sampling from a graph provides the same distribution as samples from the graphon representing the graph.

Two graphons $w_1, w_2: [0, 1]^2 \to [0, 1]$ are weakly isomorphic if there exist two measure-preserving transformations $\sigma_1, \sigma_2: [0, 1] \to [0, 1]$ such that for almost all pairs $(x, y) \in [0, 1]^2$,

$$w_1(\sigma_1 x, \sigma_1 y) = w_2(\sigma_2 x, \sigma_2 y).$$

Weakly isomorphic graphons provide the same distributions of samples.

Lovász and Szegedy [27] proved that the closure of $(\mathcal{G}, \mathcal{T})$ can be represented by the space of graphons, up to weak isomorphism. They also showed that this space is compact, which has important consequences, for example in extremal combinatorics.

Define the cut metric on the space of graphons in the following way.

$$\delta_{\square}(w_1, w_2) = \inf_{\sigma_1, \sigma_2} \sup_{S, T} \int \int w_1(\sigma_1 x, \sigma_1 y) - w_2(\sigma_2 x, \sigma_2 y) \, dy \, dx,$$

where $\sigma_1$ and $\sigma_2$ are $[0, 1] \to [0, 1]$ measure-preserving transformations, and $S$ and $T$ are measurable subsets of $[0, 1]$. Lovász and Szegedy [27] showed the inequality

$$\forall F \in \mathcal{G}, w_1, w_2 \in W : \ t(F, w_1) - t(F, w_2) \leq |E(F)| \cdot \delta_{\square}(w_1, w_2).$$
There is a much more difficult inequation about the other direction, and these together imply that the topology indicated by the cut metric $\delta$ is $\mathcal{T}$. In other words, the sequence of graphs $G_n$ is convergent if and only if this is a Cauchy-sequence with respect to $\delta$.

Summarizing, we embedded the space of all graphs into a nice and usable compact metric space, which expresses the similarity of graphs according to the sampling. Therefore, a parameter is estimable if and only if it extends to the space of graphons continuously. This space contains only graphs and limits of convergent graph sequences, therefore, if a continuous extension exists, then this is unique.

To show the power of this theory by an example, we can say that a graph is quasirandom if and only if it is close to a constant graphon in the cut metric. In other words, a graph is close to the constant $p$ graphon if and only if its sample distribution is close to the sample distribution from an Erdős–Rényi random graph with parameter $p$.

Consider now the sparse graphs. In graphs large enough, the sampling provides pairwise disjoint neighborhoods with probability tending to 1. Therefore, we modify the sampling method to the following simpler and asymptotically equivalent form. For constants $r$ and $n$, we consider the distribution of (radius) $r$-neighborhoods of a uniform random node, and we take $n$ random elements from this distribution. We call it a sample.

There are two kinds of limit objects for bounded degree graphs, both have advantages and disadvantages. One is the graphings, introduced by Elek [14] and Aldous and Lyons [1]. A graphing is given by a finite set of measure-preserving bijections on a measure space. The other limit object is the random rooted graphs. This latter one fits better to our purposes.

Which real graph parameters can be estimated by sampling is a central question of this theory. Some examples for these parameters are the number of the nodes in the largest independent set, or dominating set, or the size of the maximum matching; or the smallest number of edges that should be deleted to make the graph planar, or to separate the graphs into components of sizes at most half of the original size, all of them normalized by the number of nodes. Formally, graph parameter is a function $p: \mathcal{G} \rightarrow \mathbb{R}$, and estimator is a function mapping from samples to real numbers. We say that a parameter $p$ is estimable if for all $\varepsilon > 0$, there exists an estimator such that for all graphs $G$, the output of the estimator on a random sample from $G$ is at a distance at most $\varepsilon$ from $p(G)$ in expectation.

The estimability of a parameter expresses that the parameter is determined by its neighborhood distribution. Let us see this formally. Consider the space of all distributions of finite or infinite size bounded-degree graphs, equipped with the sigma-algebra generated by the (discrete) distributions on constant-radius neighborhoods. Let us call them random rooted graphs. For all finite graphs $G$, we assign the random rooted graph $H(G)$ as follows. We choose a node uniformly at random, and we take its component with this root. We say that a sequence of random rooted graphs is convergent if for all radius $r$, the (finite dimensional) distribution of the $r$-neighborhoods of the root converge. Denote this topological space by $X$. It is not hard to see that $p$ is estimable if and only if there exists a continuous real function $\hat{p}: X \rightarrow \mathbb{R}$ on the topological space that extends $p$, that is,

$$\forall G \in \mathcal{G}: \quad p(G) = \hat{p}(H(G)). \tag{1}$$

Now we are ready to show that some of the parameters mentioned are not testable. Consider first the smallest number of edges that should be deleted to separate the graphs into components of sizes at most half of the original size, divided by the number of nodes. For a $d$-regular expander graph sequence, this ratio tends to a positive number. However, if for each graph, we take the disjoint union of two copies of it, then this ratio is 0, because the graph already has two components of half of the original size. But the neighborhood distribution of the two sequences tend to the same random rooted graph: the $d$-regular infinite tree (as a distribution
concentrated on this only graph). Therefore, testability would require \( \tilde{p} \) at the \( d \)-regular infinite tree to be 0 and that positive number at the same time, which is a contradiction.

Another important and less obvious example is the size of the maximum independent set, divided by the number of nodes. On any \( d \)-regular bipartite graph on \( 2n \) nodes, this expected ratio is \( 1/2 \). On the other hand, on a random \( d \)-regular graph on \( 2n \) nodes, this ratio tends to less than \( 6/13 \) if \( n \to \infty \), as shown by Béla Bollobás \[5\]. As these two graph sequences tend to the \( d \)-regular infinite tree, as well, therefore, the independent ratio is not testable either. However, many other parameters, such as the relative size of the maximum matching is testable. Furthermore, the relative size of the independent set is testable for some special classes of graphs, including planar graphs.

In fact, \( \tilde{p} \) remains true even if \( \tilde{p} \) is defined only on the closure of the set of distributions corresponding to graphs \( cl\{ H(G): G \in \mathcal{G} \} \), which is a much smaller subspace. For example, if the degree of the root of a random rooted graph is 3 with probability 1, but all of its neighbors have degree 4, then this cannot be obtained as the limit of random rooted graphs \( H(G_n) \) assigned to finite graphs \( G_n \).

Denote the degree bound by \( d \). Consider a random rooted graph, and change the root to each of its neighbors with probability \( 1/d \), and with the remaining probability, keep the root at the original node. This provides another random rooted graph. If the two random rooted graphs are the same (in distribution and up to isomorphism), then we say that the random rooted graph is \textit{unimodular}. The space of unimodal random rooted graphs is denoted by \( X_u \).

The random rooted graphs assigned to finite graphs are unimodular. By the conjecture of David Aldous and Russell Lyons, the other direction is also true for the closure, that is, \( X_u = cl\{ H(G): G \in \mathcal{G} \} \). Or equivalently, for all unimodular random rooted graphs \( U \), there exists a sequence of graphs \( G_n \) such that the corresponding random rooted graphs \( H(G_n) \) tend to \( U \). This conjecture is already known in some special cases, e.g. when the distribution is concentrated on trees. \[8\] \[15\]

It turned out that the Aldous–Lyons Conjecture is strongly related to other important topics, as well. There are a number of conjectures for all countable discrete groups which are proven only for sofic groups. Mikhail Gromov asked whether all countable discrete groups are sofic. This was conjectured to be false, but there was no counterexample. Later, Gábor Elek showed, using the Cayley-graphs of the groups, that a version of the Aldous–Lyons Conjecture would imply the positive answer for the question of Gromov, which, of course, would imply the positive answer for all those conjectures for all countable discrete groups.

Therefore, and also independently of this, it would be useful to describe the space of all unimodular random rooted graphs \( X_u \), and the closure of the space of the random rooted graphs obtained from graphs \( cl\{ H(G): G \in \mathcal{G} \} \). Unfortunately, this is hopeless, because these subspaces have no nice description. Namely, in Chapter ??, we will show that some natural questions about the shape of them are algorithmically undecidable. But we also mention that the answers to all these questions are the same for the two sets, in accordance with the conjecture.

In Chapter ??, we will show that if the Aldous–Lyons Conjecture is false, then there exists a unimodular random rooted graph that, with high probability, can be distinguished from the finite graphs by the constant-radius neighborhood of only one random node. As the tool for this proof, we show that an approximately maximum flow can be constructed by a deterministic local algorithm on bounded degree graphs. Local algorithm is a concept strongly related to parameter estimation, and defined in the next subsection.
Local algorithms

A distributed algorithm on bounded degree graphs means the following. We place a processor at each vertex of the input graph, and two processors can directly communicate if they are at neighboring nodes. At the end, each processor makes some decision, and this is the output of the algorithm. For example, if we want to find a large matching, then at the end, each processor decides which of its neighbors to match with, or whether to keep unmatched. Of course, these decisions should be consistent. Distributed algorithms can be defined in several nonequivalent ways.

A local algorithm is a distributed algorithm that runs in a constant number of synchronous communication rounds, independently of the number of nodes in the network. An equivalent definition of local algorithms is that the output of each node is a function of (the isomorphism type of) the constant-radius neighborhood of the node.

Research on local algorithms was pioneered by Angluin [3], Linial [25], and Naor and Stockmeyer [30]. Angluin [3] studied the limitations of anonymous networks without any unique identifiers. Linial [25] proved some negative results for the case where each node has a unique identifier. Naor and Stockmeyer [30] presented the first nontrivial positive results. For more about local algorithms, see the recent survey paper by Suomela [32].

Randomness is a powerful and classical technique in the design of distributed algorithms, and particularly useful in breaking the symmetry [2,22,28]. For example, on transitive graphs, any local algorithm should choose the same output at each node; therefore, it is impossible to choose a positive fraction of independent vertices by a deterministic local algorithm, but this is possible with randomization. An equivalent description of random local algorithms is the following. We assign independent random seeds to the nodes, and the output at each node depends only on the constant-radius neighborhood of it, including the random seeds assigned to the vertices in the neighborhood. For example, if we choose the nodes which have a higher seed than all their neighbors, then we get an independent set of expected relative size at least $1/(d+1)$.

For typical problems, we expect from local algorithms approximate solutions only. For example, we say that we can find an almost maximum independent set if for each $\varepsilon > 0$, there exists a local algorithm that for each graph $G$, outputs an independent set, and the expected size of this set is at most $\varepsilon n$ less than the size of the maximum independent set in $G$.

Local algorithms are strongly related to parameter estimation, because many of the examined parameters come from maximization problems. For example, the size of the maximum matching, the size of the maximum independent set, or the size of the maximum cut, normalized by the number of nodes. The connection is that if we have a random local algorithm which provides an almost optimal structure, e.g. an almost maximum matching, then the relative size of the maximum matching is an estimable parameter. The estimator is the following. We take the same radius neighborhoods of a constant number of random nodes, as the radius that the random local algorithm uses. For each neighborhood, we assign random seeds to the vertices and then we calculate whether the algorithm would match the root. Then the half of the ratio of the matched nodes gives a good approximation for the relative size of the maximum matching. Huy Ngoc Nguyen and Krzysztof Onak [31] proved the estimability of several problems, e.g. the relative size of the maximum matching, by creating local algorithms.

A local algorithm with preprocessing means that the output of each node is a function of (the isomorphism type of) the graph and the constant-radius neighborhood of the node. Or equivalently, each vertex receives the same „central information“ depending on the entire graph and then they make a constant number of synchronous communication rounds, and then they present the output. This can also be interpreted as a service, as follows. There is a center with arbitrary information about the entire graph. Taking the maximum independent set problem as an example, each vertex can anonymously ask the center whether it is in the set, and the
center should answer using its preprocessed information from the graph and (the isomorphism type of) the constant-radius neighborhood of the node. These answers must be consistent, namely no two neighboring nodes should receive „yes”, but the proportion of nodes receiving „yes” should be close to the relative size of the maximum independent set.

Gábor Elek [14] proved the estimability of several parameters on graphs of subexponential growth, using a random local algorithm with the following preprocessing. He made a finite statistics of constant-radius neighborhoods of random nodes. In other words, the output at each vertex could depend on its constant-radius neighborhood and this statistics. The point of this concept is that the existence of such an algorithm giving good approximation still implies the estimability, as follows. We use the constant number of the same radius neighborhoods to make the statistics, we give this statistics to one further neighborhood, and we calculate the decision at the root. We repeat this procedure a constant number of times, from which we can get an estimation for the parameter. This observation was used for a tool to convert some statements in Borel graph theory to theorems in the field of local algorithms [17].

In Chapter ??, we compare the strengths of the different generalizations of local algorithms. In particular, we show that preprocessing is useless. More precisely, if there exists a local algorithm using preprocessing, then there exists another local algorithm with the same radius, in which the only „preprocessing” is a random variable with a continuous distribution, and this provides an output with at most the same error from the optimum, in expectation. The use of this random variable can also be interpreted as follows. We draw a local algorithm from a given probability distribution of local algorithms, and we use this at each node.

Returning to the original motivation of the sampling method, there is an experience from physicists that everything is local in all real-life graphs as well as in random graphs. While this statement is very vague, physicists assume such statements for their results whenever they can use them, and they are satisfied with the results. We show an example for a result in Chapter ??, where we give an explanation for a phase transition, using an assumption about locality. The mathematical results provided this assumption coincide with the simulation results on random graphs.

We try to find a true mathematical statement expressing the experience that „everything is local” on random graphs. In addition, this could open the door to describe the terms „typical” or „real-life” graphs better: a graph is as much „typical” as true that „everything is local” on it. There are many statements about graphs which are not true for all graphs, but which are true for the typical graphs. About such statements, all what we can do is proving that this is true for uniform random graphs on a large vertex set, with probability tending to 1. The theoretical imperfectness of this technique is pointed out by computer programs that are able to distinguish between uniform random graphs and different kind of real-life graphs, with high probability. Therefore, something to be true for almost all graphs does not mean that it is true for the typical graphs. The curiosity of the result in Chapter ?? is it proves something for typical graphs, but in an unusual sense.

We are still in the progress of describing the experience mathematically. In order to find the right statement, the best we can do is to elaborate on the easiest special cases.

The easiest problem of this kind is about the independence ratio of the 3-regular large girth graphs. First, this is a simple case because its neighborhood statistics is concentrated into the 3-regular tree. Second, the independence ratio is not determined by the neighborhood statistics. Third, it is clear what we mean random graph in this case, because the uniform random 3-regular graph has this neighborhood distribution with arbitrary small error, with probability tending to 1.

The expected relative size of the independent set generated by a local algorithm depends only on the statistics of the constant-radius neighborhoods of the graph (see Lemma ??). Furthermore, for a given neighborhood statistics, the supremum of this relative size by local
algorithms is a lower bound for the independence ratio of graphs with this statistics. The experiences shown above suggest that we can construct an almost maximal independent set on a random graph. This conjecture can be split into two parts. One is that given the neighborhood statistics, the random graph has the lowest relative size of the maximum independent set. The other is that the lowest ratio is approximable by a local algorithm. Both statements would be somewhat surprising; therefore, this question is a good indicator of whether we are on a good track to understand the experiences shown above. In Chapter ??, we show a local algorithm providing the highest independence ratio achieved so far on 3-regular large girth graphs.

There is a natural limit of local algorithms, called factor of i.i.d. (independent identical distribution) processes. This means that we assign the random seeds to the vertices, and the „local” decision should be a measurable function of the rooted graph, in a natural sense. In this language, our previous question is about the largest independence ratio achievable on 3-regular infinite trees by factor of i.i.d. processes.

While approximately maximum matching is achievable by local algorithm, it is meaningful to ask whether there exists a factor of i.i.d. maximum matching. This question is also related to group theoretic questions, through their Cayley-graphs. To show another motivation, in the proof of Miklós Laczkovich [24] for Tarski’s circle-squaring problem, the main idea was to find a perfect matching in a graphing defined by a finite number of translations. But it is still an open question whether the equidecomposability of the square and circle can be achieved with Lebesgue-measurable pieces. A similar construction with factor of i.i.d. perfect matching could provide a measurable equidecomposition.

Russell Lyons and Fedor Nazarov [29] proved that in every bipartite Cayley-graph of every non-amenable group, there is a factor of i.i.d. perfect matching. The nonbipartite case is much more difficult, but we prove in Chapter ?? that there is a factor of i.i.d. perfect matching in all nonamenable Cayley-graphs, or, in fact, in all nonamenable vertex-transitive unimodular random graphs.

Chapters ??, ?? and ?? are results of the author, based on the papers [11], [12] and [13], respectively. Chapter ?? is a joint work with Balázs Gerencsér, Viktor Harangi and Bálint Virág, Chapter ?? is a joint work with Gábor Lippner [10]; Chapter ?? is a joint work with Márton Pósfai and Yang-Yu Liu [26], and Chapter ?? is a joint work mainly with Tamás Hubai and László Lovász, but also with Omar Antolín Camarena and Gábor Lippner [9].

Hivatkozások


